

# Misspecification in mixed-model based association analysis

Willem Kruijer<sup>1</sup>

1 Biometris, Wageningen University and Research Centre, Wageningen, Netherlands

Submitted to *Genetics*

**Running head:** Misspecification in association analysis

**Key words:** misspecification, epistasis, non-additive genetic variance, missing heritability

**\*Corresponding author:**

Willem Kruijer

Biometris

Wageningen University and Research Centre

PO Box 100, 6700AC Wageningen

The Netherlands

Phone: +31 317 480806

Email: willem.kruijer@wur.nl

**Abstract:** Additive genetic variance in natural populations is commonly estimated using mixed models, in which the covariance of the genetic effects is modeled by a genetic similarity matrix derived from a dense set of markers. An important but usually implicit assumption is that the presence of any non-additive genetic effect only increases the residual variance, and does not affect estimates of additive genetic variance. Here we show that this is only true for panels of unrelated individuals. In case there is genetic relatedness, the combination of population structure and epistatic interactions can lead to inflated estimates of additive genetic variance.

Mixed models with random genetic effects have become an important tool for studying the genetic architecture of complex traits. The covariance of the genetic effects is assumed to be proportional to a genetic similarity matrix (GSM) based on a dense set of markers, which is equivalent to assuming additive effects for each standardized marker score. Under several additional assumptions, such as constant LD, this gives unbiased estimates of additive genetic variance and narrow-sense heritability ([1], [2], [3]). The sampling variance of such heritability estimators has been studied in [4] and [5]. These results are however derived under the assumption that the model is correct, i.e. contains the true distribution of the data. Here we consider situations where this is not the case, and argue that potential sources of bias may be identified by computing the parameter value  $\tilde{\theta}$  which minimizes the Kullback-Leibler divergence  $KL(Q, P_{\theta}) = \int \log(Q/P_{\theta})dQ$  with respect to the true distribution  $Q$ . It is a well known fact from statistics that in case of misspecification, i.e. when  $Q$  is not contained in the model  $\{P_{\theta} : \theta \in \Theta\}$ , the maximum likelihood (ML) estimator converges to  $\tilde{\theta}$  ([6], [7]). Several authors have studied missing or phantom heritability resulting from undetected epistatic interactions between specific loci ([8], [9], [10]). Here we investigate misspecification in a mixed model context, the covariance of the data being misspecified due to infinitesimal interactions or other non-additive effects. We consider three different scenarios (A-C), each time assuming that the additive and non-additive genetic variance is respectively 0.4 and 0.2. The total phenotypic variance is assumed to be known and equal to 1, giving a narrow- and broad-sense heritability of 0.4 and 0.6.

**Scenario A:** the phenotype  $Y = (Y_1, \dots, Y_n)'$  of  $n$  individuals is modeled using the multivariate normal distribution

$$P_{\sigma_A^2, \sigma_E^2} = N(0, \sigma_A^2 K + \sigma_E^2 I_n), \quad (1)$$

where  $K$  is a marker-based GSM,  $I_n$  the identity matrix,  $\sigma_A^2 \in [0, 1]$  is the additive genetic variance and  $\sigma_E^2 = 1 - \sigma_A^2$  is the residual variance. We assume however that  $Q$ , the *actual* distribution of  $Y$ , is the zero mean normal distribution with covariance  $0.4K + 0.2(K \cdot K) + 0.4I_n$ ,  $\cdot$  being the Hadamard (entry-wise) product. The 'epistatic' matrix  $(K \cdot K)$  is the covariance due to small epistatic interactions between all standardized marker scores (File S1). Since  $(K \cdot K)$  does not equal the identity matrix  $I_n$ ,  $Q$  is not contained in model 1. Hence, the ML-estimator will not converge to  $Q$ , but rather to the point  $(\tilde{\sigma}_A^2, \tilde{\sigma}_E^2)$  minimizing the KL-divergence  $KL(Q, P_{\sigma_A^2, \sigma_E^2})$ . For genetic similarity matrices derived from published data in maize, rice and Arabidopsis,  $\tilde{\sigma}_A^2$  ranges between 0.47 and 0.53 (Table 1). Hence, the presence of epistatic interactions leads to inflated estimates of additive genetic variance. For a panel of simulated unrelated individuals,  $\tilde{\sigma}_A^2$  is 0.40, which is due to the much smaller off-diagonal elements of  $K$ , making  $K \cdot K$  almost indistinguishable from  $I_n$ .

**Scenario B:** a plant trait is phenotyped on  $r$  genetically identical replicates. Following [5], the observations  $Y = (Y_{11}, \dots, Y_{nr})'$  are modeled by the normal distribution

$$P_{\sigma_A^2, \sigma_E^2} = N(0, \sigma_A^2 ZKZ' + \sigma_E^2 I_{nr}), \quad (2)$$

$Z$  being an incidence matrix assigning plants to genotypes. The true distribution  $Q$  is multivariate normal with covariance  $0.4ZKZ' + 0.2ZZ' + 0.4I_{nr}$ , i.e. there are non-additive (not necessarily epistatic) effects with independent  $N(0, 0.2)$  distributions. Such effects could be due to for example genotype-environment interaction. In contrast to model 1 (where  $Z = I_n$  and  $r = 1$ ),  $ZZ'$  is different from  $I_{nr}$ , and  $Q$  is not contained in model 2. Again, the value  $\tilde{\sigma}_A^2$  minimizing KL-divergence is substantially larger

than 0.4 (Table 1), and additive genetic variance will tend to be overestimated. Intuitively, this is because the block structure  $ZZ'$  is better captured by  $ZKZ'$  than by the diagonal residual.

**Scenario C** is a combination of A and B. To avoid the misspecification occurring in scenario B, the model

$$P_{\sigma_A^2, \sigma_G^2, \sigma_E^2} = N(0, \sigma_A^2 ZKZ' + \sigma_G^2 ZZ' + \sigma_E^2 I_N) \quad (3)$$

is considered, extending (2) with independent non-additive effects. This model has been used in the analysis of field trials ([11], [12]), as well as genomic prediction ([13], [14], [15]). If in fact the non-additive effects have covariance  $K \cdot K$  (as in scenario A), the data have covariance  $0.4ZKZ' + 0.2Z(K \cdot K)Z' + 0.4I_{nr}$ . As in scenarios A and B, the  $\tilde{\sigma}_A^2$  minimizing KL-divergence is larger than (Table 1), while  $\tilde{\sigma}_E^2$  was always 0.40.

Population / source	species	size ( $n$ )	A	B	C
Swedish regmap	<i>A. thaliana</i>	298	0.53	0.58	0.53
Hapmap	<i>A. thaliana</i>	350	0.47	0.60	0.48
Van Heerwaarden <i>et al.</i>	<i>Z. mays</i>	400	0.50	0.58	0.50
Zhao <i>et al.</i>	<i>O. sativa</i>	413	0.51	0.52	0.50
Unrelated individuals	simulated	3000	0.40		

Table 1: **Values of the additive genetic variance ( $\tilde{\sigma}_A^2$ ) minimizing the Kullback-Leibler divergence  $KL(Q, P)$  with respect to the true distribution ( $Q$ ) of scenarios A-C, with  $P$  contained in models 1-3.** Minimization was performed by evaluating KL-divergence on the grid  $0, 0.01, \dots, 1$  for all variance components, under the constraint they sum to one. Five populations were considered: the Arabidopsis Hapmap and Swedish regmap ([16], [5]), the rice population from [17], the maize population of [18] and a simulated population (File S1). Except for the latter, there are  $r = 2$  replicates of each genotype.

In addition to the analysis of KL-divergence we analyzed simulated traits for the first 4 populations, for which we found similar or even larger bias (File S2). This has important implications, in particular for immortal populations, for which genetically identical replicates are available (e.g. *A. thaliana*, agronomic crops, bacteria and fungi). Typically there is strong population structure and often only several hundreds of different genotypes are phenotyped. One can analyze such data at individual level (model 2) or at the level of genotypic means (model 1, with  $\sigma_E^2$  divided by the number of replicates). [5] showed that in the latter type of analysis, standard errors of heritability estimates can be huge, and recommended model 2 for both heritability estimation and genomic prediction. Here we have shown that in presence of non-additive effects, this model is likely to overestimate additive genetic variance. If however the non-additive effects are due to epistatic interactions, analysis at genotypic means level (model 1) will (apart from the large sampling variance) also give inflated estimates of additive genetic variance. This is a rather realistic scenario, since epistasis may be an important part of the genetic architecture ([19]), and several other types of non-additive effects can be ruled out or minimized for immortal populations: e.g. genotype by environment interactions are unlikely in homogeneous controlled environments with adequate randomization, and dominance effects are impossible when using inbred lines.

Interestingly, the inflation of additive genetic variance is not due to any non-linearity or absence of main effects, but rather the population structure present in the epistatic GSM, which to some extent resembles the structure of the GSM for the additive effects. At the same time, it is this structure that makes the epistatic GSM distinguishable from the diagonal error. This suggests that epistatic interactions are easier to model in structured populations, i.e. sampling variance of epistatic variance components may not be as large as in unstructured human populations ([20]). Expressions for the asymptotic variance in a model with both additive and epistatic effects (File S3) indicate that this is indeed the case. More generally, the inflation of heritability estimates due to misspecification illustrates the difficulty of mod-

eling and estimating genetic effects. As recently pointed out by [3] this is already challenging for the additive genetic effects, in the sense that depending on the genetic architecture different GSMs may be appropriate. Indeed, the potential bias resulting from an inappropriate GSM could be assessed by evaluating KL-divergence with respect to the true model, as is the case for alternatives for the epistatic GSM considered here.

## Acknowledgments

Martin Boer and Fred van Eeuwijk are acknowledged for useful comments on the manuscript. The research leading to these results has been conducted as part of the DROPS project which received funding from the European Community's Seventh Framework Programme (FP7/ 2007-2013) under the grant agreement number 244374. The research was also funded by the Learning from Nature project of the Dutch Technology Foundation (STW), which is part of the Netherlands Organisation for Scientific Research (NWO).

## Literature Cited

- [1] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42: 565–569.
- [2] Speed D, Hemani G, Johnson MR, Balding DJ (2012) Improved Heritability Estimation from Genome-wide SNPs. *The American Journal of Human Genetics* 91: 1011–1021.
- [3] Speed D, Balding DJ (2015) Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics* 16: 33–44.
- [4] Visscher PM, Goddard ME (2014) A general unified framework to assess the sampling variance of heritability estimates using pedigree or marker-based relationships. *Genetics* .
- [5] Kruijer W, Boer MP, Malosetti M, Flood PJ, Engel B, et al. (2015) Marker-based estimation of heritability in immortal populations. *Genetics* 199: 379-398.
- [6] Huber PJ (1967). The behavior of maximum likelihood estimates under nonstandard conditions. URL <http://projecteuclid.org/euclid.bsmmsp/1200512988>.
- [7] White H (1982) Maximum likelihood estimation of misspecified models. *Econometrica* 50: pp. 1-25.
- [8] Culverhouse R, Suarez BK, Lin J, Reich T (2002) A Perspective on Epistasis: Limits of Models Displaying No Main Effect. *The American Journal of Human Genetics* 70: 461–471.
- [9] Song YS, Wang F, Slatkin M (2010) General epistatic models of the risk of complex diseases. *Genetics* 186: 1467-1473.
- [10] Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* 109: 1193–1198.
- [11] Oakey H, Verbyla A, Pitchford W, Cullis B, Kuchel H (2006) Joint modeling of additive and non-additive genetic line effects in single field trials. *Theoretical and Applied Genetics* 113: 809-819.
- [12] Oakey H, Verbyla A, Cullis B, Wei X, Pitchford W (2007) Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. *TAG Theoretical and Applied Genetics* 114: 1319-1332.

- [13] Gianola D, van Kaam JBCHM (2008) Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178: 2289-2303.
- [14] Howard R, Carriquiry AL, Beavis WD (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3: Genes|Genomes|Genetics* .
- [15] Jarquin D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, et al. (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics* 127: 595-607.
- [16] Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, et al. (2012) Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat Genet* 44: 212–216.
- [17] Zhao K, Tung CWW, Eizenga GC, Wright MH, Ali ML, et al. (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nature communications* 2: 467+.
- [18] van Heerwaarden J, Hufford MB, Ross-Ibarra J (2012) Historical genomics of north american maize. *Proceedings of the National Academy of Sciences* .
- [19] Mackay TF (2014) Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature reviews Genetics* 15: 22–33.
- [20] Yang J, Lee SH, Goddard ME, Visscher PM (2011) Gcta: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* 88: 76 - 82.